Math Geosci (2019) 51:401–417
https://doi.org/10.1007/s11004-019-09795-8

Bayesian Estimation of Earth's Undiscovered Mineralogical Diversity Using Noninformative Priors

Grethe Hystad[1] · **Ahmed Eleish**[2] · **Robert M. Hazen**[3] · **Shaunna M. Morrison**[3] · **Robert T. Downs**[4]

Received: 16 February 2018 / Accepted: 11 March 2019 / Published online: 18 March 2019

Abstract Recently, statistical distributions have been explored to provide estimates of the mineralogical diversity of Earth, and Earth-like planets. In this paper, a Bayesian approach is introduced to estimate Earth's undiscovered mineralogical diversity. Samples are generated from a posterior distribution of the model parameters using Markov chain Monte Carlo simulations such that estimates and inference are directly obtained. It was previously shown that the mineral species frequency distribution conforms to a generalized inverse Gauss–Poisson (GIGP) large number of rare events model. Even though the model fit was good, the population size estimate obtained by using this model was found to be unreasonably low by mineralogists. In this paper, several zero-truncated, mixed Poisson distributions are fitted and compared, where the Poisson-lognormal distribution is found to provide the best fit. Subsequently, the population size estimates obtained by Bayesian methods are compared to the empirical Bayes estimates. Species accumulation curves are constructed and employed to estimate the population size as a function of sampling size. Finally, the relative abundances, and hence the occurrence probabilities of species in a random sample, are calculated numerically for all mineral species in Earth's crust using the Poisson-lognormal dis-

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11004-019-09795-8) contains supplementary material, which is available to authorized users.

✉ Grethe Hystad
ghystad@pnw.edu

[1] Mathematics, Statistics, and Computer Science, Purdue University Northwest, 2200 169th Street, Hammond, IN 46323-2094, USA

[2] Tetherless World Constellation, Department of Earth and Environmental Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

[3] Geophysical Laboratory, Carnegie Institution for Science, Washington, DC 20015, USA

[4] Department of Geosciences, University of Arizona, Tucson, AZ 85721-0077, USA

tribution. These calculations are connected and compared to the calculations obtained in a previous paper using the GIGP model for which mineralogical criteria of an Earth-like planet were given.

Keywords Bayesian statistics · Mixed Poisson distribution · Species estimation · Mineral frequency distribution · Mineral ecology

1 Introduction

Modeling of the frequency distribution of Earth's mineral kingdom represents an ongoing challenge in mineralogy. Mineralogists are searching for methods to estimate the total number of mineral species for subgroups of chemical composition (Grew et al. 2016; Hazen et al. 2015b). Hystad et al. (2015b) introduced a population model for the mineral species frequency distribution and found that the mineral species coupled with their localities conform to a large number of rare events (LNRE) distribution. LNRE models formulated in terms of a structural type of distribution allowed the estimation of Earth's undiscovered mineralogical diversity. This population model was used to estimate the total number of missing mineral species for different subgroups of minerals; carbon (Hazen et al. 2016), boron (Grew et al. 2016), cobalt (Hazen et al. 2017) as well as minerals with 15 diverse chemical elements (Hazen et al. 2015b). In Hystad et al. (2015b), the generalized inverse Gauss–Poisson distribution (GIGP) was fitted to the observed frequency spectrum for all mineral species, where the parameters were estimated by minimizing the simplified version of the multivariate chi-squared test for goodness of fit using the first 11 spectrum elements. In Hystad et al. (2017), the relative abundances were calculated numerically for all mineral species in Earth's crust, including the predicted undiscovered species. The standard error of the estimated population size was then calculated from parametric bootstrap samples obtained from a multinomial distribution with the relative abundances as the marginal distributions.

The objective of this paper is to estimate the total number of mineral species in Earth's crust using a Bayesian approach to species estimation. The posterior distribution for the number of species with mean, standard deviation, median, and 95% posterior intervals can then be directly estimated through Markov chain Monte Carlo (MCMC) simulations. Subsequently, the population size estimate obtained by Bayesian methods is compared to the empirical Bayes estimate, which is based on an approximation to the joint posterior distribution in concert with maximum likelihood estimation. The population size S is the parameter of interest while the other parameters in the distribution are called nuisance parameters. The joint posterior distribution is the probability of the parameter and the nuisance parameters given the data. It is proportional to the likelihood function and the joint prior probability distribution for the parameter and the nuisance parameters. In Bayesian statistics, samples are simulated from the joint posterior distribution of the parameter and the nuisance parameters. The parameters are not considered fixed as they are in classical statistics, but assumed to have their own probability distribution such that prior information about the parameters can be built into the model.

The mineral species frequency distribution, which records the number of localities for each mineral species, is right skewed with a heavy tail, and exhibit a large overdispersion (Hystad et al. 2015b). As the amount of variability in the data is greater than expected under a Poisson model, an F-mixed Poisson distribution, that offers more flexibility by allowing for variability in its parameter, is employed. The observed localities are assumed to follow a Poisson distribution, for which its unknown parameter itself is a random variable from a distribution F, called a mixing distribution. The F-mixed Poisson distribution results from integrating over the unknown parameter from the mixing distribution, providing a distribution with similar shape to the Poisson distribution, but permitting a larger variance.

In this paper, a variety of zero-truncated abundance distributions are fitted and compared; the Poisson-inverse Gaussian distribution (PIG), the generalized inverse Gauss–Poisson distribution (GIGP), and the Poisson-lognormal distribution. The negative binomial distribution based on the algorithm as described in Rodrigues et al. (2001) was also tested, but did not provide a good fit. The Poisson-lognormal distribution is another example of a large number of rare events distribution (LNRE) (Baayen 2001), but was not used in Hystad et al. (2015b) because of computational difficulties. Baroni and Evert (2007) also found in initial evaluations that the lognormal distribution had inferior performance in extrapolation tasks compared to the other LNRE models. The Poisson-lognormal distribution is employed in the Bayesian framework and found to provide a better prediction than the GIGP distribution used in Hystad et al. (2015b) for all mineral species. In addition, the entire frequency spectrum is fitted to the various models instead of right-truncating the sample to only the rare species. In Hystad et al. (2015b), only the first 11 spectrum elements along with the total observed number of species were taken into account. An advantage to the Bayesian approach is that it is relatively straightforward to modify the computer code for the different abundance distributions. As an addition to the paper, the code in the statistical software package, R, is provided in Online Resources 1–3. The R code is inspired by C code given by Quince et al. (2008). In Hystad et al. (2015b, 2017), the total number of distinct mineral species in Earth's crust were estimated to be 6, 394 with a standard error of 110 using the GIGP distribution to model the observed frequency spectrum. It was argued in Hystad et al. (2015b) that this number was an underestimate, and mineralogists have since found this value to be unreasonably low as about 100 new mineral species are discovered each year.

A prevailing method in species estimation is to fit several parametric species abundance distributions to the data and compare the fits with some goodness-of-fit measure. It is a universal law in ecology that every community shows a hollow curve with many rare species and a few abundant species. However, an open problem is where the predictable differences in nature and structure of the species abundance distributions occur (McGill et al. 2007). Many different species abundance distributions often fit a data set well, and it is often difficult to distinguish them since the data is usually noisy and the differences are most unequivocal in the tails, where the sampling is usually incomplete (McGill et al. 2007).

The data in the present study consist of a list of mineral species and their localities as of February 2014 from the crowd-sourced website Mindat.org. There are 135,415 distinct localities and, when counted over all mineral species, these data provide a sample size of 652,856 observations, where each observation is a unique mineral species-locality pair. As of February 2014, there were 4831 approved mineral species reported from Earth's crust, where 22% of the mineral species are found at only one locality, 12% are found at only two localities, while more than half of all mineral species are found at 5 or fewer localities.

The task of estimating the number of biological species in an ecological population has been studied since the celebrated paper by Fisher et al. (1943). An overview of research and methods to estimate the number of species is provided in Bunge and Fitzpatrick (1993), Bunge and Barger (2008), and Barger and Bunge (2010). Barger and Bunge (2008, 2010) estimate the number of species in a Bayesian framework for the mixed-Poisson likelihood by deriving expressions for two objective priors based on Jeffreys' prior (Jeffreys 1946) and reference priors (Bernardo 1979; Bernardo and Ramón 1998) from the full likelihood. The information matrix for the likelihood is derived using the linear difference score for discrete parameters as given in Lindsay and Roeder (1987). The information matrix is then a combination of discrete and continuous-valued parameters with diagonal elements that factor into a product of the function of the discrete parameter S and a function of the continuous nuisance vector θ. The objective priors then have independent components for the discrete parameter and the nuisance parameters. Reference priors were recommended over the Jeffreys' prior for multiparameter problems. Barger and Bunge (2010) found that the marginal reference prior for the discrete parameter S does not depend on the number of nuisance parameters, while the Jeffreys' prior is a function of the number of nuisance parameters.

In Barger and Bunge (2008), the Poisson distribution and the exponential-mixed Poisson distribution were employed as the abundance distributions and precise forms were given for the joint reference priors for S and the nuisance parameter. Finding an analytical reference prior for a vector-valued nuisance parameter is difficult. In Barger and Bunge (2010), the gamma, inverse Gaussian and the two-mixed exponential were used for the mixture distributions in a Poisson-mixture model, but different objective marginal priors than the reference or Jeffreys' prior were employed for the nuisance parameters. Independent priors were chosen for the nuisance parameters. For the inverse Gaussian, the multivariate reference prior for the two component nuisance parameters as derived in Gutiérrez-Peña and Rueda (2003) was used. Using the generalized inverse Gaussian (GIG) and the lognormal distribution as the mixture distributions, it can be checked that the information matrices factor into a function of S and a function of the nuisance parameters by using the method derived in the paper of Barger and Bunge (2010), but the elements of the matrices do not factor into separate functions of the nuisance parameters. Similar issues were found in Barger and Bunge (2010) for both two- and three-parameter vectors for the nuisance parameters.

In Quince et al. (2008), an improper uniform prior for S along with flat priors for the nuisance parameters were used for several different abundance distributions including the generalized inverse Gauss–Poisson, the Poisson-lognormal, and the Poisson log-Student's t distributions. The models were fit to the entire observed frequency spectra.

The lognormal distribution employed as the species abundance distribution has been used extensively in ecology (McGill et al. 2007) as well as a model for word frequency distributions (Baayen 2001; Carroll 1967). The lognormal distribution seems to provide good fits to species-rich communities (McGill et al. 2007).

2 Bayesian Approach

Let S denote the population size of distinct mineral species in Earth's crust. Assume each mineral species has a population probability π_i (relative abundance) of being sampled at an arbitrary locality, where $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_S$ defines the ordering schemes and $\sum_{i=1}^{S} \pi_i = 1$. The probability of a given mineral being found is assumed to be constant over all localities. The relative abundances are given by $\pi_i = \frac{x_i}{M}$, where x_i is the number of localities for the ith species in the population and M is the total population number (sample size for observing S mineral species). Let a sample of N mineral species-locality pairs be drawn randomly and independently from the total population of occurrences of S mineral species. The number of observed localities for the ith species in a sample of N mineral species-locality pairs then follows approximately a Poisson distribution with mean $\lambda_i = \frac{x_i}{M} N$ (Hystad et al. 2015b; Quince et al. 2008). The unknown population abundances x_i of the individual mineral species are assumed independent samples from a species abundance distribution $F(x_i|\theta')$ representing heterogeneity in the population. The distribution F is indexed with the m dimensional hyper-parameter θ' whose density is denoted by $\pi(\theta')$. The probability that an individual species will occur j times in a sample of size N is given by the F-mixed Poisson distribution

$$P_{\theta',v}(j) = \int_0^\infty \frac{e^{-xv}(xv)^j}{j!} f(x|\theta')\,\mathrm{d}x,$$

for $j = 0, 1, 2, \ldots$, where $f(x|\theta') = \frac{\partial}{\partial_x} F(x|\theta')$ and where $v = \frac{N}{M}$ is the sampling frequency (Quince et al. 2008). Thus, $P_{\theta',v}(0)$ represents the probability of not being observed. Using a change of variable $\lambda = xv$ and noticing that the three abundance distributions employed in this paper are invariant under this transformation, it follows that

$$P_\theta(j) = \int_0^\infty \frac{e^{-\lambda}\lambda^j}{j!} f(\lambda|\theta)\,\mathrm{d}\lambda,$$

where the rescaled parameter θ is a function of the sampling frequency (Quince et al. 2008).

It is known that the distribution of minerals is not random, and certainly exhibits inter-species correlations (Hystad et al. 2015b). For example, some groups of mineral species tend to be found together in the same locations. In addition, less commonly occurring minerals form in specific geologic settings and, hence, are found in mineral deposits in restrictive locations; for example, molybdenite minerals (Golden et al. 2013). These mineral species will have a higher probability of being found in these geologic settings than outside. Big, colorful crystals are also much more likely to be found than microscopic, poorly crystallized minerals, which are in turn much more likely to be found than those requiring an electron microscope to identify. Even though the Poisson distribution assumes independence and consistency of probabilities, and the data is certainly not quite independent with constant probabilities, it is still worth the effort to examine the estimate of the model.

Let n_j represent the number of mineral species observed at exactly j localities. The sample size or the number of mineral species-locality pairs is given by $N = \sum_{j\geq 1} jn_j$. The total number of observed distinct mineral species in a sample of size N is given by $w = \sum_{j\geq 1} n_j$. As of February 2014, the total number of distinct mineral species discovered is $w = 4831$ with $N = 652{,}856$ species-locality pairs. The number of distinct mineral species found at, for example, only one or two localities are $n_1 = 1062$ and $n_2 = 569$, respectively. The likelihood is

$$L(S, \theta|\text{data}) = \binom{S}{w}(1 - P_\theta(0))^w (P_\theta(0))^{S-w} \frac{w!}{\prod_{j\geq 1} n_j!} \prod_{j\geq 1} \left(\frac{P_\theta(j)}{1 - P_\theta(0)}\right)^{n_j}. \quad (1)$$

Equation (1) can be written in the form

$$L(S, \theta|\text{data}) = A(w|S, \theta)B(n_1, n_2, \ldots|\theta), \quad (2)$$

where $A(w|S, \theta)$ is a binomial likelihood for w with success probability $1 - P_\theta(0)$ and $B(n_1, n_2, \ldots|\theta)$ is a multinomial likelihood for the observed frequencies with the zero-truncated F-mixed Poisson distribution as the probabilities (Barger and Bunge 2008, 2010; Quince et al. 2008; Rodrigues et al. 2001).

The joint posterior distribution for S and the nuisance parameters θ is given by

$$P(S, \theta|\text{data}) \propto L(S, \theta|\text{data})\pi(S, \theta),$$

where we can write

$$L(S, \theta|\text{data}) = \frac{S!}{(S - w)!} \frac{1}{\prod_{j\geq 1} n_j!} (P_\theta(0))^{S-w} \prod_{j\geq 1} (P_\theta(j))^{n_j},$$

and where the joint prior distribution $\pi(S, \theta)$ can be factored as a product of the prior distribution for S and the prior distribution for θ as in $\pi(S, \theta) = \pi(S)\pi(\theta)$ (Barger and Bunge 2010).

In this paper, several choices for the abundance distribution P_θ are employed; the Poisson-lognormal and the Poisson-inverse Gaussian (PIG), which are two parameter distributions, and the generalized inverse Gauss–Poisson (GIGP) distribution, also called the Sichel distribution (Sichel 1971), which is a three-parameter distribution. Noninformative, improper uniform priors are used for S and the nuisance parameters for all distributions. The deviance information criterion (DIC) is used to compare the different fits by the Bayesian approach, and the model with the smallest DIC indicates a better fit. Let $\hat{\psi}_{\text{Bayes}}$ be the posterior median of the vector of parameters. The deviance information criterion (DIC) is given by (Gelman et al. 2013)

$$\text{DIC} = -2\log(L(\hat{\psi}_{\text{Bayes}}|\text{data})) + 2p_{\text{DIC}},$$

where P_{DIC} is the effective number of parameters given by

$$p_{\text{DIC}} = 2[\log(L(\hat{\psi}_{\text{Bayes}}|\text{data})) - E_{\text{post}}(\log(L(\psi|\text{data})))].$$

Here E_{post} is the average of ψ over its posterior distribution calculated by

$$\frac{1}{M}\sum_{i=1}^{M}\log(L(\psi_i|\text{data})),$$

where ψ_i, $i = 1, 2, \ldots, M$ are the simulated values computed from the posterior distribution using MCMC methods.

The parameterization of the Sichel distribution used for $P_\theta(j)$ in this paper is given in Stein et al. (1987). It is

$$P_\theta(j) = \frac{\left(\frac{\omega}{\alpha}\right)^{\gamma}\left(\frac{\beta\omega}{\alpha}\right)^{j} K_{j+\gamma}(\alpha)}{K_\gamma(\omega)\, j!}, \tag{3}$$

where $\omega = \sqrt{\beta^2 + \alpha^2} - \beta$, $K_\gamma(z)$ is the modified Bessel function of the second kind of order γ and argument z, and $\alpha, \beta > 0$, $-\infty < \gamma < \infty$ for frequency classes $j = 0, 1, 2, \ldots$. According to Stein et al. (1987), this particular parameterization of the Sichel distribution results in lower correlation between the maximum likelihood estimates of the parameters α and β when γ is in a neighborhood of -0.5 as compared to alternative parameterizations. Setting $\gamma = -0.5$ gives the Poisson-inverse Gaussian distribution (PIG). In Hystad et al. (2015b), the Sichel distribution was found to fit well to the mineral frequency spectrum using the parameterization given in Baayen (2001). The Poisson-lognormal distribution with parameters $-\infty < \mu < \infty$ and $\sigma > 0$ is given by

$$P_\theta(j) = \frac{1}{\sqrt{2\pi}\sigma j!}\int_0^\infty e^{-\lambda}\lambda^{j-1}e^{-\frac{(\ln(\lambda)-\mu)^2}{2\sigma^2}}\,\mathrm{d}\lambda,$$

where the integral does not have a closed form. In this paper, the uniform prior for the Poisson-lognormal distribution is compared to a simplified version of the reference prior. The elements of the information matrix for the Poisson-lognormal distribution factor into a function of S and the nuisance parameters $\theta = (\mu, \sigma)$, but not additionally into separate functions of μ and σ. Instead, a simplified version of the priors based on the two-parameter reference prior for the lognormal distribution and the reference prior, $S^{-\frac{1}{2}}$, for S, given in Barger and Bunge (2010), is employed. The Fisher information matrix for the lognormal distribution with mean μ and variance σ^2 is given by $H(\mu, \sigma^2) = Diag(\frac{1}{\sigma^2}, \frac{1}{2\sigma^4})$ with the inverse denoted by R. Using the fact that the matrix elements of R and H factor as $r_{11} = r_1(\mu)r_1(\sigma^2) = \sigma^2$ and $h_{22} = h_2(\mu)h_2(\sigma^2) = \frac{1}{2\sigma^4}$, respectively, and applying corollary 1 in Bernardo and Ramón (1998), the joint reference prior for μ and σ^2 is given by the well-known result $\pi(\mu, \sigma^2) = r_1(\mu)^{-\frac{1}{2}}h_2(\sigma^2)^{\frac{1}{2}} \propto \frac{1}{\sigma^2}$. Thus, a simplified version of the reference prior is given by $\pi(S, \mu, \sigma) \propto S^{-\frac{1}{2}}\frac{1}{\sigma}$.

3 Empirical Bayes Estimator

In the empirical Bayes approach, the approximation $\hat{p}(S|\text{data}) \sim p(S|\hat{\theta}, \text{data})$ is used, where $\hat{\theta}$ is the maximum likelihood estimator of θ (Rodrigues et al. 2001). For large values of S, the empirical Bayes estimate of S is the same as the one obtained by the Bayesian approach (Rodrigues et al. 2001). The empirical Bayes estimator can be obtained in the following way: First, $\hat{\theta}$ is computed by maximizing $B(n_1, n_2, \ldots |\theta)$ in Eq. (2) with respect to θ. This is equivalent to computing the maximum likelihood estimator $\hat{\theta}$ from $W = w$ identically distributed zero-truncated F-mixed Poisson variables. Given $\hat{\theta}$, the second maximization gives the empirical Bayes estimator $\hat{S}_{\hat{\theta}} = \left[\frac{W}{1-P_{\hat{\theta}}(0)}\right]$ where $[k]$ is the integer part of k (Chao and Bunge 2002; Rodrigues et al. 2001). Notice that if θ was known, $\hat{S}_{\theta}$ would be an unbiased estimator of S, which follows from the fact that W is a binomial random variable with parameters S and $1 - P_{\theta}(0)$.

4 Method

Samples from the joint posterior distribution for the population size S and the nuisance parameters were simulated by applying the random walk Metropolis algorithm (Chib and Greenberg 1995) using the software C and the statistical software package R. Different numbers of simulation runs were used for the different abundance distributions. MCMC simulations for the Sichel distribution and the Poisson-inverse Gaussian were run in C using code from Quince et al. (2008). The code was also modified to be run in R as given in Online Resources 2–3. For the Poisson-lognormal distribution, the code was run in R as given in Online Resource 1. In order to run the code in R, the gsl library was employed, which is an R wrapper for the special functions of the Gnu scientific library. The simulations were run with three chains from overdispersed starting values. The mixing of the chains and the convergence of the algorithm were checked using the R libraries, rstan (Stan Development Team 2017) and coda (Plummer et al. 2018). Convergence and mixing are achieved when the potential scale reduction factor, $\hat{R}$, is below 1.2, and the effective sample size is at least 10 per chain (Gelman et al. 2013). The normal distribution was used as the proposal distribution with the mean set at the current value in the algorithm, and the standard deviation was adjusted to achieve a certain acceptance rate. The models were fit to the entire observed frequency spectrum without truncating for large frequencies. For the Sichel and the Poisson-inverse Gaussian distributions taken as the abundance distributions, 10 million simulation runs were completed with the first 5 million iterations discarded. The conservative choice of discarding the first half of the iterations was made in order to reduce the impact of the starting values (Gelman et al. 2013). The distribution of the simulated parameter values is then close to the joint posterior distribution for a large enough number of iterations. The values from each chain were combined into one sample and then values from every 30,000th and 3000th iteration were selected for the Sichel and the Poisson-inverse Gaussian distributions, respectively, to reduce autocorrelation. The standard deviation for the proposal distribution was set such that an acceptance rate of about 44% and 41% was reached for the Sichel and the Poisson-inverse Gaussian distributions, respectively. For the Poisson-lognormal distribution with a noninformative, improper uniform prior and a reference prior, the acceptance rates were 45%, where 1 million simulation runs from each chain were completed and combined with the first half of the iterations discarded. Here, values from every 500th iteration were selected.

The median, the mean with a standard error, and standard deviation were calculated, and 95% posterior intervals were estimated by the 2.5% and 97.5% order statistics. The different models were compared using the deviance information criteria (DIC), and the model with the smallest DIC is preferred. It is difficult to determine what constitutes a significant difference in DIC. According to Spiegelhalter et al. (2002), a difference in DIC values between five and ten is substantial, while a difference of more than ten indicates that the model with the higher DIC may be ruled out. Quince et al. (2008) considered a difference in DIC of less than or equal to six not significantly different if the models had the same number of parameters.

The empirical Bayes estimates calculated for the various abundance distributions are compared to the estimates obtained by the Bayesian approach. The zero-truncated distributions, SI and SICHEL, which are different parameterizations of the Sichel distribution, and the zero-truncated distribution, Poisson-inverse Gaussian (PIG), all from the R library gamlss.tr (Stasinopoulos and Rigby 2018) are fitted to the observed frequency spectrum through maximum likelihood methods. In addition, the zero-truncated Poisson-lognormal distribution from the R library poilog (Grøtan and Engen 2015) is fitted. For the Poisson-lognormal distribution, the standard error is computed from 5000 parametric bootstrap samples that are simulated using the estimated parameters.

5 Results

The median, mean, standard error (SE) of the mean, empirical standard deviation (SD), 95% posterior interval, and the deviance information criterion (DIC) for the population size S calculated from samples simulated by MCMC methods using different abundance distributions with noninformative uniform priors are provided in Table 1. A simplified version of the reference prior is also used for the Poisson-lognormal distribution. The standard error (SE) of the mean is calculated using the effective sample size that takes into account the autocorrelation in the sample. However, since the samples were thinned by selecting every kth element from the simulations, the time series standard errors of the mean were close to the standard errors of the mean based on independent samples. The empirical Bayes estimates for the population size S, where the abundance distribution is taken to be the Poisson-lognormal, Poisson-inverse Gaussian, and two different parameterizations of the Sichel distribution are given in Table 2. Notice that the two parameterizations of the Sichel distribution given in Table 2 are different than the one used for the Bayesian approach.

The median is close to the mean for all distributions. The Poisson-lognormal distribution is the preferred model for the abundance distribution followed by the Sichel distribution for all mineral species since it has the smallest DIC. The DIC for the

Table 1 The median, mean, standard error (SE) of the mean, empirical standard deviation (SD), 95% posterior interval, and the deviance information criterion (DIC) for the population size S using Bayesian methods

Abundance distribution	Median	Mean	SE	SD	95% Posterior interval	DIC
Poisson-lognormal	9308	9322	8	363	(8650, 10,070)	2875.7
Poisson-lognormal with reference prior	9287	9309	8	360	(8665, 10,082)	2875.7
Sichel (GIGP)	6198	6194	7	96	(6009, 6370)	2882.9
Poisson-inverse Gaussian (PIG)	6080	6080	1	72	(5950, 6230)	2887.7

Table 2 Bayes estimate for the population size S for various abundance distributions. The standard error (SE) is included for the Poisson-lognormal distribution

Abundance distribution	Estimate of S
Poisson-lognormal empirical Bayes	9285 (341)
SI empirical Bayes	6212
SICHEL empirical Bayes	6215
PIG empirical Bayes	6085

Poisson-inverse Gaussian is significantly larger than for the Poisson-lognormal distribution and will therefore give a worse prediction. The estimate of S is calculated as the median of the posterior sample. Using the Poisson-lognormal abundance distribution, the total number of distinct mineral species, S, in the Earth's crust is estimated to be 9308. Thus, 4477 mineral species are yet to be found. A 95% posterior interval for S is estimated to be (8650, 10,070). In comparison, using the Sichel distribution, the estimate of S is 6198 and a 95% posterior interval for S is given to be (6009, 6370). An estimate of 9308 mineral species seems to be reasonable. Hazen et al. (2015a) estimate that at least 15,300 plausible mineral species can form on terrestrial planets and moons. However, many of these species require combinations of chemical and physical conditions that are unlikely, and thus, they will occur on a small fraction on Earth-like planets. However, many of these species are unlikely to form (Hazen 2017). Here, we estimate that approximately 9300 of these mineral species will occur on Earth at any given time. The empirical Bayes estimates are close to the estimated values of S obtained by the Bayesian methods. This indicates that for large values of the population size S, mineralogists can use the empirical Bayes estimate for S for subgroups of mineral species. The empirical Bayes method has the advantage that it is much simpler to use than fully Bayesian methods, which require more advanced statistical computing. However, Rodrigues et al. (2001) found in their paper that the empirical Bayes estimate is not accurate for small population values.

Figure 1 shows the posterior distributions of S that were obtained from the MCMC simulations using the Poisson-lognormal distribution (Fig. 1a) and the Sichel distribution (Fig. 1b) as the abundance distributions. The expected number of mineral species from j number of localities is given by $S * P_\theta(j)$. Figure 2a–c shows the bar plot of the observed frequencies with the expected frequencies drawn as points for the Poisson-lognormal distribution (Fig. 2a), the Sichel distribution (Fig. 2b), and the

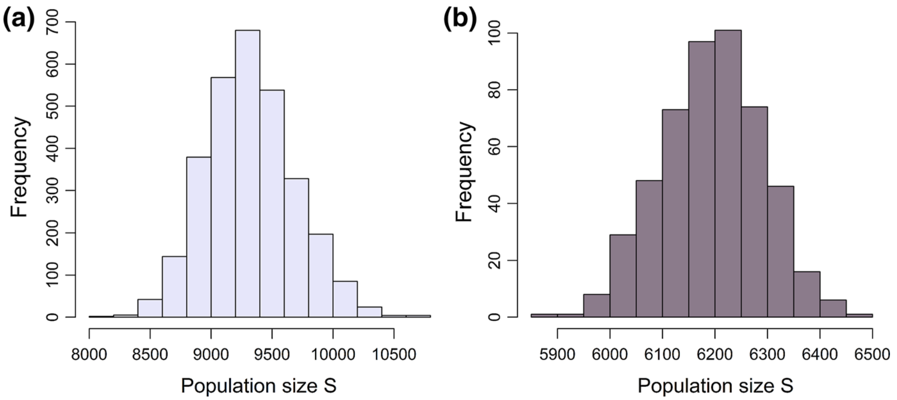

Fig. 1 Posterior distribution for the population size S for **a** the Poisson-lognormal distribution and for **b** the Sichel distribution computed by MCMC methods

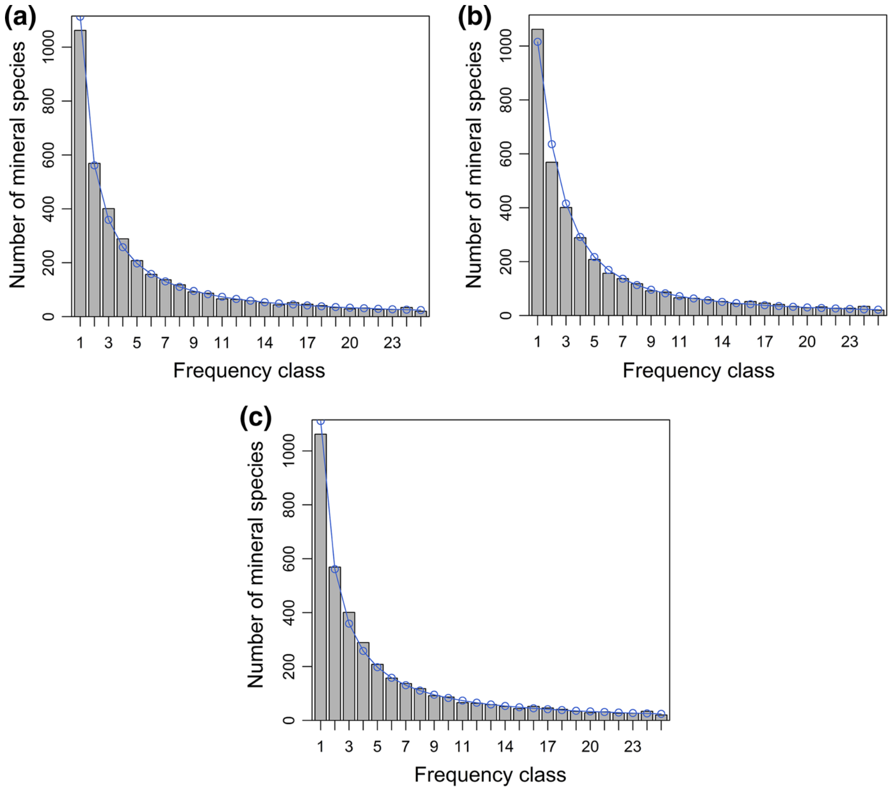

Fig. 2 Observed and expected number of mineral species for a given frequency class for **a** Poisson-lognormal (Bayesian methods), for **b** Sichel (Bayesian methods), and for **c** Poisson-lognormal distribution (empirical Bayes estimator)

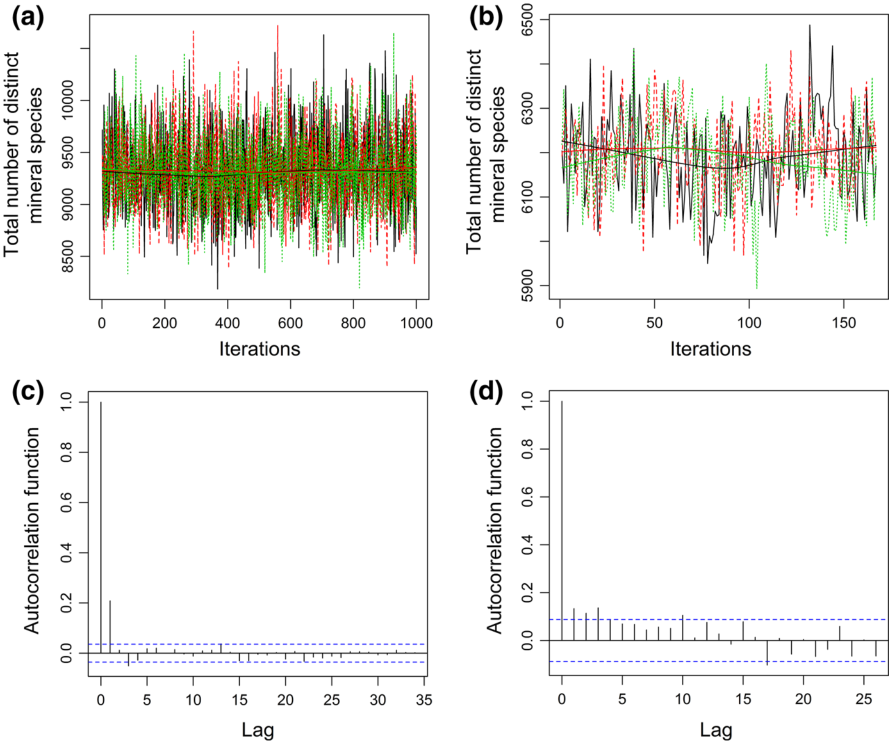

Fig. 3 Trace plot for the population size S for **a** the Poisson-lognormal distribution and for **b** the Sichel distribution. Autocorrelation plot for the population size S for **c** the Poisson-lognormal distribution and for **d** the Sichel distribution

Bayes estimate using the Poisson-lognormal distribution (Fig. 2c). The predicted frequencies from the Bayes estimate using the Poisson-lognormal distribution is given by $w * \frac{P_\theta(j)}{1-P_\theta(0)}$, where $\frac{P_\theta(j)}{1-P_\theta(0)}$ is the zero-truncated Poisson-lognormal density function. We see from the graphs that these abundance distributions have a good fit to the data for all mineral species. Figure 3 shows the trace plot for three chains with smooth lines going through the plot for the Poisson-lognormal distribution (Fig. 3a) and the Sichel distribution (Fig. 3b), and autocorrelation plots of S using the Poisson-lognormal distribution (Fig. 3c) and the Sichel distribution (Fig. 3d). Autocorrelation is reduced by selecting only every kth element from the simulations. The trace plots also indicate achieved convergence.

5.1 Mineral Species Accumulation Curves

In this section, the expected mineral species accumulation curves for the Poisson-lognormal and the Sichel distribution are computed. The accumulation curves allow the prediction of total number of mineral species as a function of sample size. Using the change of variable $\lambda = xv$ as described in Sect. 2 for the Poisson-lognormal distribution, the mean of the lognormal distribution is transformed while the standard deviation stays fixed. Thus, the lognormal distribution has parameters $\mu = \mu' + \ln(v)$ and σ, where μ' and σ are the mean and standard deviation, respectively, of the log transformed abundance. If the sample size is changed to N_{new} with corresponding sampling frequency, $v_{\text{new}} = \frac{N_{\text{new}}}{M}$, the new mean is given by

$$\mu_{\text{new}} = \mu' + \ln(v_{\text{new}}) = \mu + \ln\left(\frac{N_{\text{new}}}{N}\right).$$

Thus, for new sample sizes, the probability density function can be computed in terms of the original fitted parameters (Quince et al. 2008). The expected number of observed species is then coming from a binomial distribution with mean

$$S(1 - P_{\theta_{\text{new}}, v_{\text{new}}}(0)), \tag{4}$$

where $P_{\theta_{\text{new}}, v_{\text{new}}}(0)$ is a function of the rescaled parameter $\theta_{\text{new}} = (\mu_{\text{new}}, \sigma)$ (Quince et al. 2008).

The mixture distribution corresponding to the Sichel distribution given by Stein et al. (1987) and in Eq. (3) is the generalized inverse Gauss distribution (GIG) given by the parameterization in Jørgensen (1982). It can be checked that the GIG distribution can be written in the form

$$f(\lambda|\theta') = \frac{\eta'^{-\gamma}}{2K_\gamma(\omega)}\lambda^{\gamma-1}\exp\left[-\frac{1}{2}\omega(\eta'\lambda^{-1} + \eta'^{-1}\lambda)\right],$$

with $\omega = \sqrt{a'^2 + b'^2} - b'$, $\eta_{\prime} = \frac{b'}{N}$, and $a' = \omega\sqrt{1 + N\frac{2\eta'}{\omega}}$, and where a', b', and γ are treated as the parameters. Using the transformation $\lambda = xv$, the rescaled parameters are $\eta = \eta' v$, and hence, $b = b'v$ and $a = \omega\sqrt{1 + N\frac{2\eta}{\omega}}$ with fixed parameter γ. If the sample size is changed to N_{new} with corresponding sampling frequency $v_{\text{new}} = \frac{N_{\text{new}}}{M}$, the new values are

$$\begin{aligned}
\eta_{\text{new}} &= \eta' v_{\text{new}} = \eta\frac{N_{\text{new}}}{N},\\
a_{\text{new}} &= \omega\sqrt{1 + N_{\text{new}}\frac{2\eta_{\text{new}}}{\omega}} = \omega\sqrt{1 + \frac{2\eta}{\omega}\frac{(N_{\text{new}})^2}{N}}\\
&= \omega\sqrt{1 + \left(\frac{N_{\text{new}}}{N}\right)^2\left(\frac{a^2 - \omega^2}{\omega^2}\right)},
\end{aligned}$$

and

$$b_{\text{new}} = \eta\frac{(N_{\text{new}})^2}{N} = b\left(\frac{N_{\text{new}}}{N}\right)^2.$$

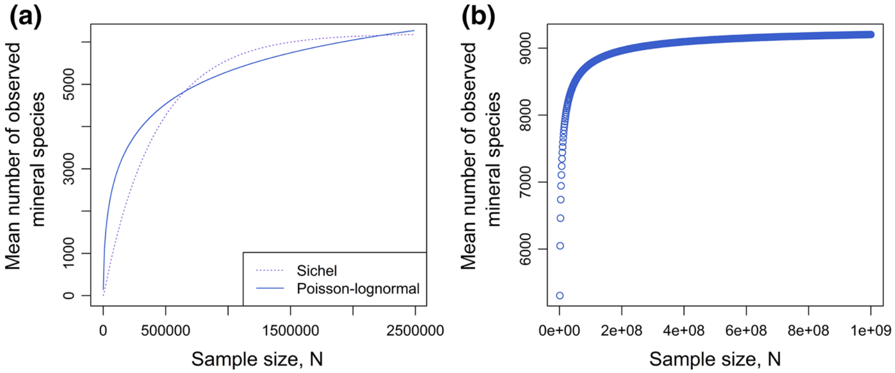

Fig. 4 Expected number of observed mineral species for **a** the Poisson-lognormal and the Sichel distribution for sample sizes up to 2.5 millions and for **b** the Poisson-lognormal distribution for sample sizes between 1 million and 1 billion

It is not necessary to compute b_{new} directly for the accumulation curve as ω stays fixed under the transformation. The expected number of observed species using the Sichel distribution is then given by Eq. (4) with parameters $\theta_{\text{new}} = (a_{\text{new}}, b_{\text{new}}, \gamma)$. Figure 4 shows the expected mineral species accumulation curves for both the Poisson-lognormal distribution and the Sichel distribution (Fig. 4a) and for the Poisson-lognormal distribution for large sampling sizes (Fig. 4b).

5.2 Occurrence Probabilities

In Hystad et al. (2017), the relative abundances for all mineral species in Earth's crust were calculated using the Sichel distribution. A similar calculation is performed in this paper for the Poisson-lognormal distribution. The details can be found in Hystad et al. (2017). The value of σ is independent of the sampling frequency, while the value of μ changes. The value of μ for complete sampling for the lognormal distribution, can be found by solving the equation $9308 = S = \exp(\frac{1}{2}\sigma^2 - \mu)$ (Baayen 2001). This value is used in the calculation for the relative abundances. Let the number of mineral species in the population with probability greater than or equal to ρ be approximated by the continuous function

$$G(\rho) = \int_{\rho}^{1} g(\pi)\, \mathrm{d}\pi,$$

where $g(\pi) = \frac{f(\pi)}{\pi}$ and where $f(\pi)$ is the lognormal distribution. The equations of interest are $G(\rho_k) = k$ for $k = 1, 2, \ldots, 9308 = S$, which are solved by employing the Newton–Raphson method. The population probabilities are then found by

$$\pi_1 = \int_{\rho_1}^{1} f(\pi)\, \mathrm{d}\pi,$$

$$\pi_k = \int_{\rho_k}^{\rho_{k-1}} f(\pi)\, \mathrm{d}\pi,$$

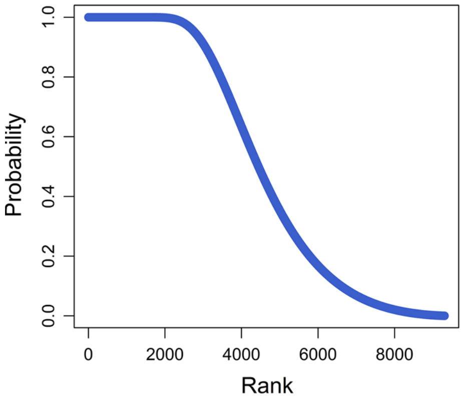

Fig. 5 The probabilities that mineral species are present at least one time in a random sample of $N = 652{,}856$ mineral species-localities pairs from the crust of the Earth versus population rank using the Poisson-lognormal distribution

for $k = 1, 2, \dots, 9308 = S$ (Hystad et al. 2017). Using the relative abundances, π_k, the occurrence probabilities for each of the 9308 mineral species in a random sample of size $N = 652{,}856$ mineral species-locality pairs from Earth's crust are calculated from the binomial distribution. The results are illustrated in Fig. 5, where the species with rank 1 is the most abundant species in the population. A similar graph using the Sichel distribution was applied in Hystad et al. (2017) to obtain mineralogical criteria to quantify an Earth-like planet. Rounded to three decimal places, there are 1973 mineral species with greater than 0.999 probability of occurring on all Earth-like planets, whereas 2774 mineral species have at least 0.950 probability of occurrence. These are the rock-forming minerals as described in Hazen et al. (2015a) and Hystad et al. (2015a), which are a necessity for an Earth-like planet. Similar numbers were found in Hystad et al. (2017). However, the Poisson-lognormal distribution predicts several more species than the Sichel distribution. Thus, there are 186 species that have less than 0.0010 probability of occurring. The low-probability species are the ones that occur by chance.

By generating two random samples of size $N = 652{,}856$ mineral species-locality pairs from the multinomial distribution with the relative abundances as the marginal distributions, the expected number of species that will be different in the two samples was estimated. Using the Poisson-lognormal distribution, it was found that about 16% of species are expected to be different in two random samples of size N from two modeled Earth-like planets.

6 Conclusions

The zero-truncated Poisson-lognormal distribution is found to provide the best fit to the mineral frequency distribution using Bayesian methods. Employing this distribution, the total number of mineral species in Earth's crust was estimated to be 9308. The empirical Bayes estimates for various abundance distributions were found to be close to the estimates obtained by Bayesian methods. The relative abundances and the occurrence probabilities were calculated numerically for all mineral species in Earth's crust including the predicted undiscovered species. Subsequently, a replaying of mineral evolution on Earth, or similarly, the comparison of two Earth-like planets would result in a variety of 16% different mineral species distributed over the two samples.

It would be interesting to compute and compare the reference priors for the various distributions with two and three nuisance parameters used in this paper by applying the method in Barger and Bunge (2010). In a future paper, a model that reflects age, geographic, spatial correlation effects, tectonic factors, and other restrictive factors will be incorporated as covariates for the abundance distribution. New mineral species are being discovered more frequently than in the past as a result of new and improved high-resolution sampling techniques. The age of discovery is therefore an important covariate to be included in a future model. To incorporate covariates into the model, a generalized additive model for location, scale, and shape (GAMLSS) (Stasinopoulos and Rigby 2018) may be used. In this model, the distribution parameters can be functions of the explanatory variables with a monotonic link function relating the distribution parameters to the predictors. Several sub-models of the GAMLSS exists, including a parametric linear GAMLSS model with no additive terms in the distribution parameters, and a semi-parametric additive GAMLSS, where the additive term is a smooth function of the explanatory variables.

Acknowledgements We would like to thank the reviewers for their valuable comments to improve the paper. We gratefully acknowledge support from NASA Mars Science Laboratory Mission NNX11AP82A, as well as support from the Alfred P. Sloan Foundation (Grant No. 2013-10-01), the W.M. Keck Foundation (Grant No. 140002372), the John Templeton Foundation (Grant No. 60645), the Deep Carbon Observatory, the Carnegie Institution for Science, and an anonymous private foundation.

References

Baayen RH (2001) Word frequency distributions, text, speech and language technology, vol 18. Kluwer Academic Publishers, Dordrecht

Barger K, Bunge J (2008) Bayesian estimation of the number of species using noninformative priors. Biom J 50(6):1064–1076

Barger K, Bunge J (2010) Objective bayesian estimation for the number of species. Bayesian Anal 5(4):765–785

Baroni M, Evert S (2007) Words and echoes: assessing and mitigating the non-randomness problem in word frequency distribution modeling. In: Proceedings of the 45th annual meeting of the association for computational linguistics, Prague, Czech Republic, pp 904–911

Bernardo JM (1979) Reference posterior distributions for bayesian inference. J R Stat Soc B 41:113–147

Bernardo JM, Ramón JM (1998) An introduction to bayesian reference analysis: inference on the ratio of multinomial parameters. J R Stat Soc D 47:101–135

Bunge J, Barger K (2008) Parametric models for estimating the number of classes. Biom J 50(6):971–982

Bunge J, Fitzpatrick M (1993) Estimating the number of species: a review. J Am Stat Assoc 88(421):364–373

Carroll JB (1967) On sampling from a lognormal model of word frequency distribution. In: Kučera H, Francis WN (eds) Computational analysis of present-day American English. Brown University Press, Providence, pp 406–424

Chao A, Bunge J (2002) Estimating the number of species in a stochastic abundance model. Biometrics 58(3):531–539

Chib S, Greenberg E (1995) Understanding the metropolis-hastings algorithm. Am Stat 49(4):327–335

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol 12(1):42–58

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis, 3rd edn. CRC Press, Boca Raton

Golden J, McMillan M, Downs RT, Hystad G, Goldstein I, Stein HJ, Zimmerman A, Sverjensky DA, Armstrong JT, Hazen RM (2013) Rhenium variations in molybdenite (MoS_2): evidence for progressive subsurface oxidation. Earth Planet Sci Lett 366:1–5

Grew ES, Krivovichev SV, Hazen RM, Hystad G (2016) Evolution of structural complexity in boron minerals. Can Mineral 54:125–143

Grøtan V, Engen S (2015) Poisson lognormal and bivariate Poisson lognormal distribution, package poilog. https://cran.r-project.org/web/packages/poilog/poilog.pdf. Accessed 20 Feb 2015

Gutiérrez-Peña E, Rueda R (2003) Reference priors for exponential families. J Stat Plan Inference 110:35–54

Hazen RM (2017) Chance, necessity, and the origins of life: a physical sciences perspective. Philos Trans R Soc A 375:20160353

Hazen RM, Grew ES, Downs RT, Golden J, Hystad G (2015a) Mineral ecology: chance and necessity in the mineral diversity of terrestrial planets. Can Mineral 53:295–324

Hazen RM, Hystad G, Downs RT, Golden JJ, Pires AJ, Grew ES (2015b) Earth's 'missing' minerals. Am Mineral 100:2344–2347

Hazen RM, Hummer DR, Hystad G, Downs RT, Golden JJ (2016) Carbon mineral ecology: predicting the undiscovered minerals of carbon. Am Mineral 101:889–906

Hazen RM, Hystad G, Golden JJ, Hummer DR, Liu C, Downs RT, Morrison SM, Ralph J, Grew ES (2017) Cobalt mineral ecology. Am Mineral 102:108–116

Hystad G, Downs RT, Grew ES, Hazen RM (2015a) Statistical analysis of mineral diversity and distribution: earth's mineralogy is unique. Earth Planet Sci Lett 426:154–157

Hystad G, Downs RT, Hazen RM (2015b) Mineral species fequency distribution conforms to a large number of rare events model: prediction of Earth's missing minerals. Math Geosci 47:647–661

Hystad G, Downs RT, Hazen RM, Golden JJ (2017) Relative abundances of mineral species: a statistical measure to characterize earth-like planets based on earth's mineralogy. Math Geosci 49:179–194

Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proc R Soc Ser A 186:453–461

Jørgensen B (1982) Statistical properties of the generalized inverse Gaussian distribution, 1st edn. Springer, New York

Lindsay BG, Roeder K (1987) A unified treatment of integer parameter models. J Am Stat Assoc 82:758–764

McGill BJ et al (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol Lett 10:995–1015

Plummer M, Best N, Cowles K, Vines K, Sarkar D, Bates D, Almond R, Magnusson A (2018) Output analysis and diagnostics for MCMC, package 'coda'. https://cran.r-project.org/web/packages/coda/coda.pdf. Accessed 8 Oct 2018

Quince C, Curtis TP, Sloan WT (2008) The rational exploration of microbial diversity. Int Soc Microb Ecol J 2:997–1006

Rodrigues J, Milan LA, Leite JG (2001) Hierarchical Bayesian estimation for the number of species. Biom J 43(6):737–746

Sichel HS (1971) On a family of discrete distributions particularly suited to represent long-tailed frequency data. In: Proceedings of the third symposium on mathematical statistics, Pretoria, South Africa, pp 51–97

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc B 64:583–639

Stan Development Team (2017) RStan: the R interface to stan. http://mc-stan.org/, R package version 2.17.2

Stasinopoulos M, Rigby B (2018) Generalised additive models for location scale and shape, package GAMLSS. https://cran.r-project.org/web/packages/gamlss/gamlss.pdf. Accessed 6 Oct 2018

Stein GZ, Zucchini W, Juritz JM (1987) Parameter estimation for the sichel distribution and its multivariate extension. J Am Stat Assoc 82:938–944